Gafurov, T.; Usaola, J.; Prodanovic, M. (2015). Incorporating spatial correlation into stochastic generation of solar radiation data. *Solar Energy*, v. 115, pp. 74-84.
DOI: 10.1016/j.solener.2015.02.018

Incorporating spatial correlation into stochastic generation of solar radiation data

Tokhir Gafurov [a,*], Julio Usaola [b], Milan Prodanovic [a]

[a] *Electrical Systems Unit, IMDEA Energy Institute, Avenida de Ramón de la Sagra 3, 28935 Móstoles, Madrid, Spain*
[b] *Department of Electrical Engineering, Universidad Carlos III de Madrid, Avenida de la Universidad 30, 28911 Leganés, Madrid, Spain*

Communicated by: Associate Editor Frank Vignola

Abstract

Spatial correlation of solar radiation (SCSR) has a significant impact on the overall data quality when generating radiation time series for multiple sites. Currently, there are no known methods for integration of SCSR into synthetic data by using reduced and easily avail able inputs. Based on a hypothesis that at long timescales general and simple characterization of SCSR is possible, this paper addresses the problem of modeling monthly and daily SCSR. A regression analysis of satellite derived radiation data covering over 300,000 loca tions pairs in 4 US regions is firstly described and general mathematical expressions for SCSR estimation are presented. A procedure for incorporating spatial correlation into conventional stochastic solar radiation models is then introduced by applying the obtained SCSR formulae and the existing methods of linear algebra. Finally, the underlying hypothesis is validated and the effectiveness of the proposed technique for creating spatially correlated monthly and daily solar radiation values is demonstrated based on numerical simulations and analysis of historical data.

Keywords: Regression analysis; Solar resource; Spatial correlation; Synthetic generation

1. Introduction

It is a common practice to use synthetic solar radiation data when meteorological measurements for a certain location or timescale are unavailable or unreliable. The synthetic data in this case can be generated by multivariate and univariate statistical models. The former, also known as the weather generator, creates the radiation time series at long timescales (usually on daily basis) together with other weather parameters (Wilks and Wilby, 1999). The univariate model generates only the values of solar radiation at various timescales based on simple inputs, typically comprising the long-term statistics for clearness index (Badescu, 2008; Meteonorm, 2013; Watgen, 1992). Despite high performance these stochastic algorithms traditionally have not incorporated spatial correlation of solar radiation (SCSR), which has a significant impact on the overall data quality when synthesizing the radiation time series for multiple sites.

The review of the recent literature indicates that there have not been any successful attempts to integrate SCSR into univariate solar radiation models. Yet, a number of methods for multivariate algorithms have been proposed. The effective and relatively simple approach in this case seems to be not to change the conventional single-site weather model, but to modify the input random number

* Corresponding author.
E mail address: tokhir.gafurov@imdea.org (T. Gafurov).

Nomenclature

A, B	sub- and superscripts referring to different locations
C	constant in the Haversine formula
C_r	correlation matrix for random number streams
d	intersite distance (km)
K_t	clearness index
$K_{t,M}$	long-term average of the monthly clearness index
n_p	polynomial degree
r, r_{corr}	uncorrelated and correlated random numbers
R^2	coefficient of determination
U	upper triangular matrix in the Cholesky factorization
x_j	mathematical indicator for intersite dependence (MIID)
y, y'	actual and predicted values of a given variable
ΔK_t	first difference (ramp rate) of clearness index series
λ	longitude
ϕ	latitude
v_i	constants in the functional relation between C_r and PCC
MAE	mean absolute error
MARE	mean absolute relative error
NSDD	normalized SDD
PCC	Pearson's correlation coefficient for ΔK_t^A and ΔK_t^B
RMSE	root mean square error
SDD	standard deviation of difference $\Delta K_t^A - \Delta K_t^B$

streams instead, for the given network of locations so that the resultant synthetic values of solar radiation (or other weather parameter) have realistic spatial correlations (Wilks, 1999; Khalili et al., 2009). Even though this technique allows adequate reproduction of SCSR, it still requires synchronized regional meteorological measurements for tuning the model coefficients, which, to some extent, defeats the purpose of using the synthetic data.

Apparently, there are no known methods for multisite generation of solar radiation data from reduced (easily available) inputs. This might be explained by the absence of general formal description of SCSR that could allow estimating the spatial correlation for any two locations based on simple predictor variables. Obviously, at short timescales, since local weather factors become significant, such a general characterization of SCSR is unrealistic. However, at long timescales the authors believe that this is possible. Interestingly, the literature reveals few research papers addressing this subject. For example, Aguado (1986) and Suckling (1995) estimate the coefficients of variability (the standard deviation of the intersite daily radiation differences divided by the mean values) for the selected station pairs by using directly the measured daily solar radiation values. And in recent studies, Hoff and Perez (2012) and Badosa et al. (2013) evaluate SCSR by Pearson's correlation coefficient and they represent the solar resource via ramp rates (deltas) of clear-sky index, which is more suitable for such analyses (Badescu, 2008). The paper by Hoff and Perez (2012) is of particular interest as it employs satellite-derived data covering over 70,000 pairs of points, though the considered timescales are only up to 4 h. The common shortcoming of the existing studies is that they try to relate SCSR to the intersite distance only, whereas the results clearly show that this dependence changes from one region to the other.

Taking into account the mentioned gaps in current research, the aims of this work are to: (a) determine general mathematical expressions relating solar radiation between two sites at monthly and daily timescales; and (b) incorporate SCSR into existing univariate algorithms for generating solar radiation data based on the obtained expressions. The given tasks are covered separately in Sections 2 and 3, respectively. In the end of the paper the concluding remarks are provided.

2. Characterization of the spatial correlation of solar radiation

2.1. Objectives

The authors hypothesize that at long timescales a general relation for SCSR can be derived based on simple (easily accessible) inputs. From various available parameters the clearness index K_t is selected to represent solar radiation, by taking into account: (a) its common use in climate research, particularly in the area of synthetic data generation; and (b) its straightforward calculation allowing exclusion of the local factors such as site altitude and turbidity levels required in the clear-sky index estimation. In order to remove non-stationarity (trend) in the K_t time series, differencing is applied, which means that the focus is not on the actual values of K_t, but on its ramp rates ΔK_t.

The final objective in this case is defined as to perform a regression analysis of historical meteorological data and to determine mathematical expressions that would allow quantifying SCSR for any two locations at monthly and daily timescales by using intersite distance and monthly statistics of K_t.

2.2. Methods

The adopted methodology involves:

1. use of extensive meteorological data covering multiple geographical regions and climate zones to capture general patterns of SCSR,
2. simplification of descriptive models to avoid overfitting (Koller and Sahami, 1996).

The main aspects of the performed regression analysis are described in detail below. Considering heuristic nature of the study the authors admit that the selected methods and the obtained results might not be optimal.

All simulations and data analyses were done in Matlab.

2.2.1. Data

Two datasets were used in the study. The main set, taken from the satellite-derived SolarAnywhere Data (available within the US National Solar Radiation Database (NSRDB, 2013)), comprises the hourly radiation data for the period of 1998 2009 at 1591 locations evenly distributed (grid spacing 0.2°) over 4 US regions as shown in Fig. 1a. The total number of location pairs is over 300,000 with intersite distances in the range of 15 540 km (mean 205 km). The regions were selected arbitrarily to represent various climates and thus to increase the data diversity. The boundaries and resolution of the data grid points had to be limited to reduce the computational time. The chosen distance range was found to be sufficient to model fully SCSR at a daily timescale.

The additional dataset, taken from ASDC (2013), consists of daily solar radiation values for the period of 1993 2004 at 84 locations (grid spacing 0.5°) primarily over Spain and Germany as shown in Fig. 1b. The total number of location pairs is over 1500 with intersite distances in the range of 65 550 km (mean 295 km). The additional set was used only for testing the performance of the final regression models.

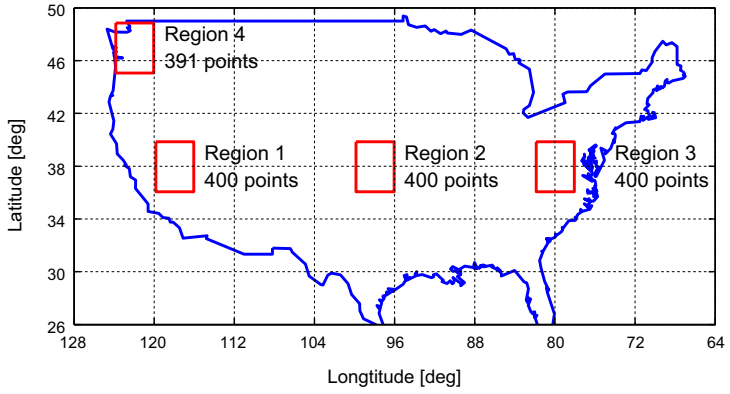
The clearness index, when not given directly, was calculated from its definition as the ratio between the global solar radiation on a horizontal plane and the corresponding extraterrestrial radiation. The latter was estimated based on formulae from Duffie and Beckman (2006).

2.2.2. Initial domain of variables

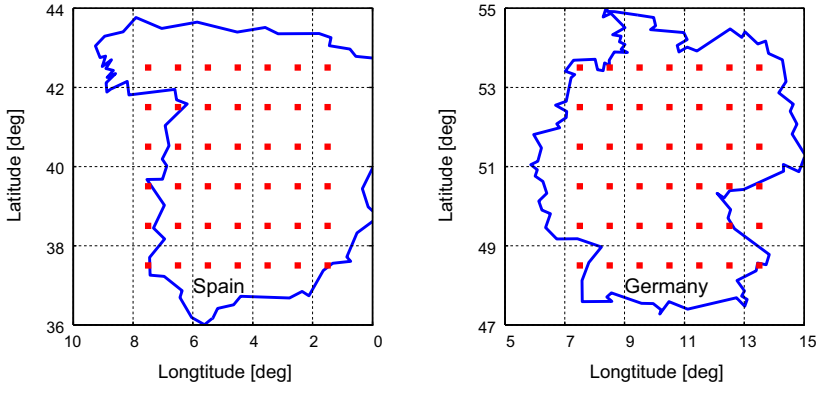
The initial domain of variables refers to the set of response (output) and explanatory (input) parameters that are used during the regression analysis. The selected candidate response variables or SCSR estimators consist of:

standard deviation of difference

$$\mathrm{SDD} = \mathtt{std}(\Delta K_t^A - \Delta K_t^B) \tag{1}$$



(a) Main locations in the US



(b) Additional locations in Europe

Fig. 1. Selected sites for solar radiation data.

normalized standard deviation of the difference

$$\mathrm{NSDD} = \frac{\mathrm{SDD}}{\left[\mathtt{std}(\Delta K_t^A)\,\mathtt{std}(\Delta K_t^B)\right]^{0.5}} \tag{2}$$

Pearson's correlation coefficient

$$\mathrm{PCC} = \mathtt{corr}(\Delta K_t^A, \Delta K_t^B) \tag{3}$$

where the superscripts A and B refer to the two locations in the data pair. Note, that PCC can alternatively be expressed as:

$$\mathrm{PCC} = \frac{\mathtt{std}(\Delta K_t^A)^2 + \mathtt{std}(\Delta K_t^B)^2 - \mathrm{SDD}^2}{2\,\mathtt{std}(\Delta K_t^A)\,\mathtt{std}(\Delta K_t^B)} \tag{4}$$

In the case of explanatory parameters, it was assumed that besides the distance d, SCSR can also be characterized by certain mathematical indicators of intersite dependence (MIID) x_j determined by the monthly average values of the clearness index $K_{t,M}$. The authors chose arbitrarily 16 MIIDs as shown in Table 1. Thus, the initial set of input variables was limited to d and x_{1-16}.

The intersite distance was calculated based on the Haversine formula (Sinnott, 1984):

$$\begin{cases} C = \sin^2\frac{\phi_A-\phi_B}{2} + \sin^2\frac{\lambda_A-\lambda_B}{2} \times \cos\phi_A \times \cos\phi_B \\ d = 2 \times 6371 \times \mathrm{atan2}(\sqrt{C}, \sqrt{1-C}), \end{cases} \tag{5}$$

where ϕ and λ are the latitude and longitude of the site.

2.2.3. Functional form of the regression

The final data fitting was done with linear and quadratic polynomials by using the Matlab functions `polyfitn` and `polyvaln` from D'Errico (2011). The intercept was not forced to zero (or unity), because it reduces the model performance to a certain extent. In order to avoid overfitting, the number of explanatory variables was limited to two, which means that, in addition to the intersite distance d, the authors had to choose one of the candidate MIIDs defined in Table 1.

The polynomials of a higher degree ($n_p > 2$), not preferred due to their complexity (large number of terms), were still tested in the preliminary studies, which showed that the regression model would not improve significantly and it would usually produce unrealistic nonlinear patterns.

Table 1
Chosen mathematical indicators of intersite dependence.

x_1	$\{\mathtt{mean}(K_{t,M}^A) + \mathtt{mean}(K_{t,M}^B)\}/2$
x_2	$\{\mathtt{max}(\mathtt{K}_{\mathtt{t,M}}^{\mathtt{A}}) \quad \mathtt{min}(\mathtt{K}_{\mathtt{t,M}}^{\mathtt{A}}) + \mathtt{max}(\mathtt{K}_{\mathtt{t,M}}^{\mathtt{B}}) \quad \mathtt{min}(\mathtt{K}_{\mathtt{t,M}}^{\mathtt{B}})\}/2$
x_3	$\{\mathtt{std}(K_{t,M}^A) + \mathtt{std}(K_{t,M}^B)\}/2$
x_4	$\{\mathtt{mean}(\lvert\,\Delta K_{t,M}^A\,\rvert) + \mathtt{mean}(\lvert\,\Delta K_{t,M}^B\,\rvert)\}/2$
x_5	$\{\mathtt{max}(\Delta\mathtt{K}_{\mathtt{t,M}}^{\mathtt{A}}) \quad \mathtt{min}(\Delta\mathtt{K}_{\mathtt{t,M}}^{\mathtt{A}}) + \mathtt{max}(\Delta\mathtt{K}_{\mathtt{t,M}}^{\mathtt{B}}) \quad \mathtt{min}(\Delta\mathtt{K}_{\mathtt{t,M}}^{\mathtt{B}})\}/2$
x_6	$\{\mathtt{std}(\Delta K_{t,M}^A) + \mathtt{std}(\Delta K_{t,M}^B)\}/2$
x_7	$\lvert\,\mathtt{mean}(K_{t,M}^A) \quad \mathtt{mean}(K_{t,M}^B)\,\rvert$
x_8	$\lvert\,\mathtt{max}(\mathtt{K}_{\mathtt{t,M}}^{\mathtt{A}}) \quad \mathtt{min}(\mathtt{K}_{\mathtt{t,M}}^{\mathtt{A}}) \quad \mathtt{max}(\mathtt{K}_{\mathtt{t,M}}^{\mathtt{B}}) + \mathtt{min}(\mathtt{K}_{\mathtt{t,M}}^{\mathtt{B}})\,\rvert$
x_9	$\lvert\,\mathtt{std}(K_{t,M}^A) \quad \mathtt{std}(K_{t,M}^B)\,\rvert$
x_{10}	$\lvert\,\mathtt{mean}(\lvert\,\Delta K_{t,M}^A\,\rvert) \quad \mathtt{mean}(\lvert\,\Delta K_{t,M}^B\,\rvert)\,\rvert$
x_{11}	$\lvert\,\mathtt{max}(\Delta\mathtt{K}_{\mathtt{t,M}}^{\mathtt{A}}) \quad \mathtt{min}(\Delta\mathtt{K}_{\mathtt{t,M}}^{\mathtt{A}}) \quad \mathtt{max}(\Delta\mathtt{K}_{\mathtt{t,M}}^{\mathtt{B}}) + \mathtt{min}(\Delta\mathtt{K}_{\mathtt{t,M}}^{\mathtt{B}})\,\rvert$
x_{12}	$\lvert\,\mathtt{std}(\Delta K_{t,M}^A) \quad \mathtt{std}(\Delta K_{t,M}^B)\,\rvert$
x_{13}	$\mathtt{std}(K_{t,M}^A \quad K_{t,M}^B)$
x_{14}	$\mathtt{std}(\Delta K_{t,M}^A \quad \Delta K_{t,M}^B)$
x_{15}	$\mathtt{corr}(K_{t,M}^A, K_{t,M}^B)$
x_{16}	$\mathtt{corr}(\Delta K_{t,M}^A, \Delta K_{t,M}^B)$

Note: The ramp rate $\Delta K_{t,M}$ is defined as the difference between the $K_{t,M}$ values for the current and previous month.

2.2.4. Model performance assessment and variables selection

The goodness of fit was evaluated by the coefficient of determination R^2, which is a commonly reported measure of regression fit (Hansen, 2013) and a reasonable estimator of single variable relevance (Guyon and Elisseeff, 2003). R^2 equals to the square of the correlation coefficient between the actual and predicted values of a given output variable.

The parameters selection in this study was straightforward and involved data fitting at various combinations of response (SDD, NSDD or PCC) and explanatory (d, x_j) variables and choosing the most adequate ones based on the R^2 value.

The dependence of the regression fit on the number of predictors was estimated based on step-wise regression (Guyon and Elisseeff, 2003), which starts with d in the model and at each step adds the indicator x_j that improves the model performance most.

Considering a large sample size, the application of more advanced statistical tests such as the Fisher test was found to be unnecessary (Hansen, 2013; Lin et al., 2013). For the same reason the cross-validation during variables selection was not required (Guyon and Elisseeff, 2003).

In addition to the quantitative assessments, the visual examination of the model predictions and mispredictions was also conducted, though not presented in this paper.

2.3. Results

The regression analysis was performed on a workstation (Intel Xeon W3503, 12 GB RAM, 2.4 GHz) in three general steps: (a) the required data for each site were extracted from the original source files and the missing variables were calculated; (b) the selected SCSR estimators were determined for each location pair; and (c) the linear and quadratic polynomial regressions were fitted for different combinations of the response and explanatory parameters. The total computational time for the main dataset was more than 9 h.

2.3.1. Regional features of the main dataset

According to the adopted approach for the SCSR characterization, the main data should have a large number of samples and a certain level of diversity. To demonstrate the latter, the plots of PCC versus distance at monthly, daily and hourly timescales are presented in Fig. 2 for each of

the selected US regions. As one can see, there are notable variations in SCSR within and among the regions. It is clear that distance cannot be used as the only predictor.

The spatial correlation in general decreases with higher distance and lower timescale. This trend, however, is less pronounced and even reversed when moving to an hourly timescale. As shown in Fig. 2 after the initial sharp drop the hourly PCC declines with the distance very slowly and maintains relatively high values (at some point exceeding the corresponding daily PCC) even at the distances for which the hourly weather changes at two locations are expected to be nearly independent. The reason is that the clearness index probability distribution depends not only on the sky conditions (e.g. cloud cover, aerosol content), but also on the optical air mass (Badescu, 2008; Perez et al., 1990), which means that the hourly K_t variations are driven to a certain extent by the diurnal cycle of the sun.

2.3.2. Regression analysis

The main results from the step-wise regressions are presented in Fig. 3. The significance of various MIIDs when using the selected response parameters is compared in Fig. 4. These findings together with the visual data inspection show:

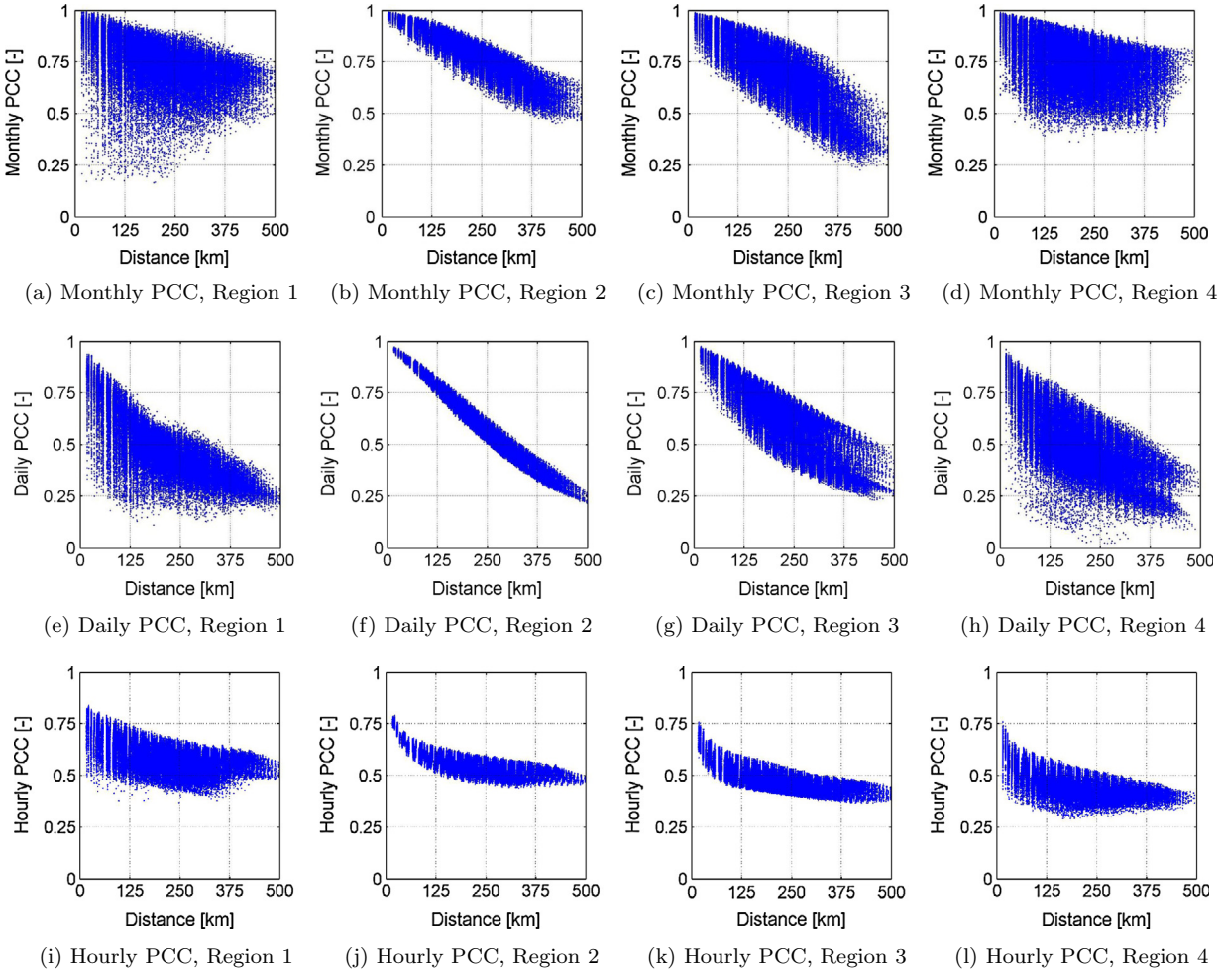
1. The adequate response variables are SDD for a monthly timescale and NSDD and PCC for a daily timescale.
2. The combination of distance with only one MIID as input improves notably the model performance ($R^2 > 0.8$). The most relevant indicator seems to be x_{14}. With the addition of further MIIDs the corresponding gain is negligible.
3. For the monthly SCSR the linear polynomial is sufficient, whereas for the daily SCSR the quadratic approximation is more accurate.

The obtained final regression functions are:

$$\mathrm{SDD}_{month} = 4.462\mathrm{e}-05\,d + 1.0594\,x_{14} + 0.017, \tag{6}$$

$$\mathrm{NSDD}_{day} = -1.0769\mathrm{e}-06\,d^2 - 0.0302\,d\,x_{14} + 0.0022\,d - 132.3022\,x_{14}^2 + 22.8342\,x_{14} + 0.2428, \tag{7}$$

$$\mathrm{PCC}_{day} = 6.6044\mathrm{e}-08\,d^2 + 0.0162\,d\,x_{14} - 0.0013\,d + 70.8353\,x_{14}^2 - 15.5606\,x_{14} + 1.0516, \tag{8}$$

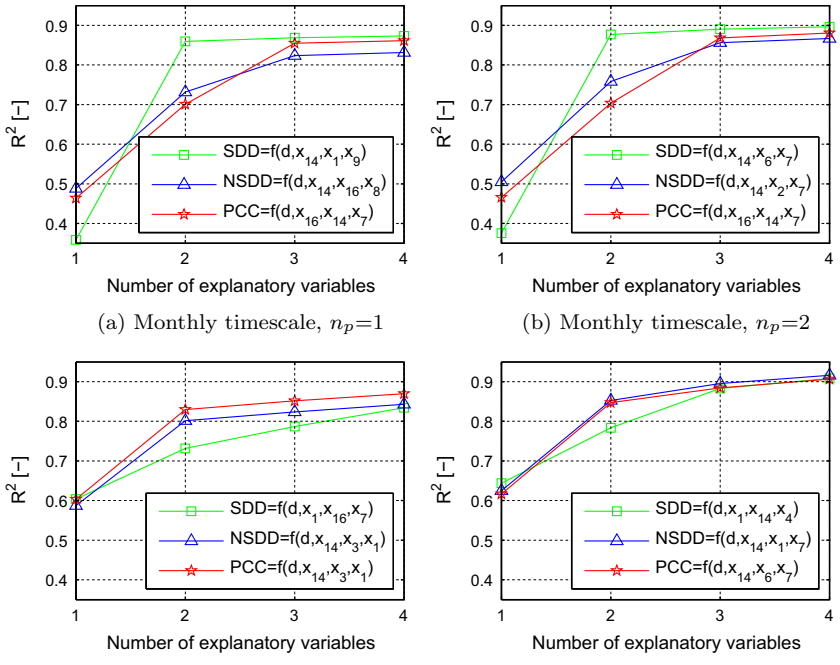


(a) Monthly PCC, Region 1 (b) Monthly PCC, Region 2 (c) Monthly PCC, Region 3 (d) Monthly PCC, Region 4

(e) Daily PCC, Region 1 (f) Daily PCC, Region 2 (g) Daily PCC, Region 3 (h) Daily PCC, Region 4

(i) Hourly PCC, Region 1 (j) Hourly PCC, Region 2 (k) Hourly PCC, Region 3 (l) Hourly PCC, Region 4

Fig. 2. Actual variation of PCC with distance for the location pairs in the selected 4 US regions.

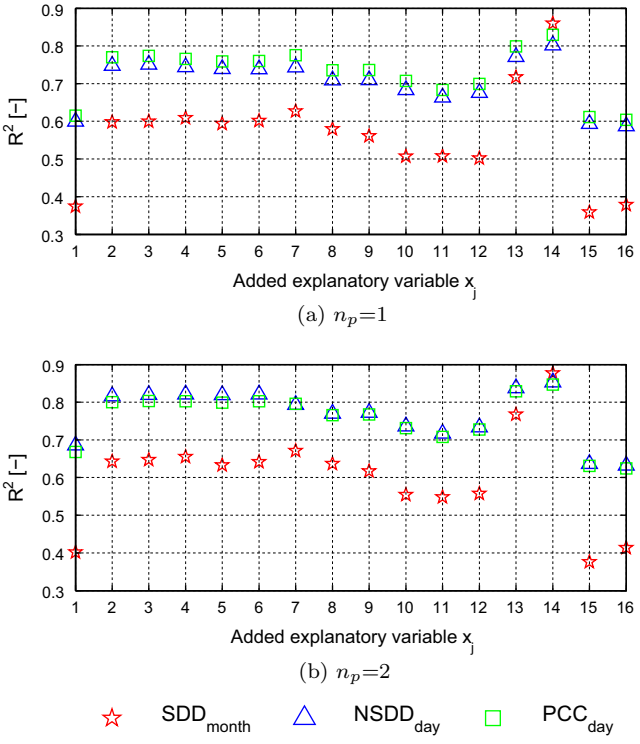


(a) Monthly timescale, n_p=1 (b) Monthly timescale, n_p=2

(c) Daily timescale, n_p=1 (d) Daily timescale, n_p=2

Note: The legend shows the identified explanatory variables in the order of aggregation.

Fig. 3. Results of the step wise regression (main dataset).



(a) n_p=1

(b) n_p=2

Fig. 4. R^2 values of the regression models using MIID as an additional explanatory variable besides distance (main dataset).

where the input parameters are limited as:

$$15 \leqslant d \leqslant 500, \; x_{14} \leqslant 0.06 \tag{9}$$

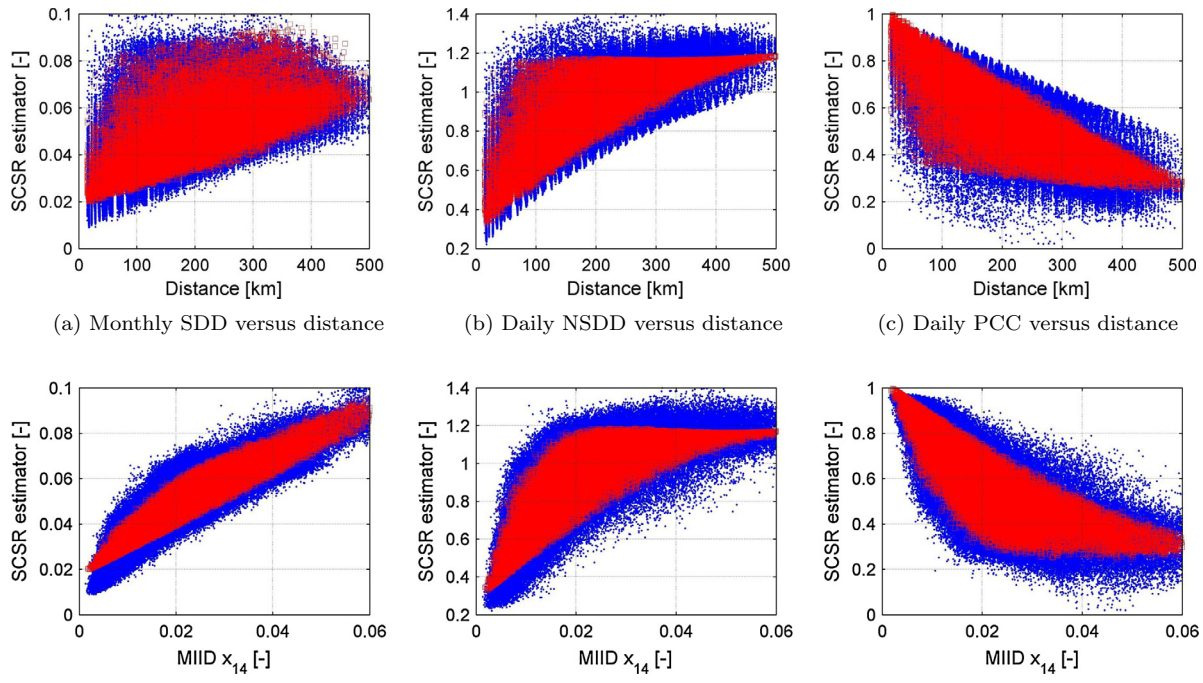
$$d \leqslant 650 - 7500 x_{14} \text{ (only for daily SCSR)} \tag{10}$$

The restriction associated with Eq. (10) approximately represents the boundary after which the values of the daily SCSR estimators stagnate; according to Eqs. (7) and (8) it roughly corresponds to $\mathrm{NSDD}_{day} = 1.2$ and $\mathrm{PCC}_{day} = 0.3$.

The real and predicted variations of the chosen SCSR estimators with d and x_{14} are illustrated in Figs. 5 and 6 for the main and additional (validation) datasets. One can observe how the regression models capture the linear or non-linear trends in the actual values of the response variables.

The goodness of fit statistics of the derived regression functions is summarized in Table 2. Besides R^2, the measures of model fit include: the root mean squared error RMSE $\sqrt{\mathrm{mean}\left((y-y')^2\right)}$, the RMSE variation coefficient CV(RMSE) $\frac{\mathrm{RMSE}}{\mathrm{mean}(y)}$, the mean absolute error MAE $\mathrm{mean}(|y-y'|)$ and the mean absolute relative error MARE $\mathrm{mean}(|y-y'|/|y|)$, where y and y' denote the real and predicted values of a given output parameter, respectively. As one can see, the deviations in the case of the additional dataset are somewhat higher, but still reasonable; this confirms the overall adequacy of the proposed regression models and it thus demonstrates the feasibility of the characterization of SCSR at long timescales. It is important to be cautious, however, when applying the given models for climate zones highly divergent from that covered by the main dataset.

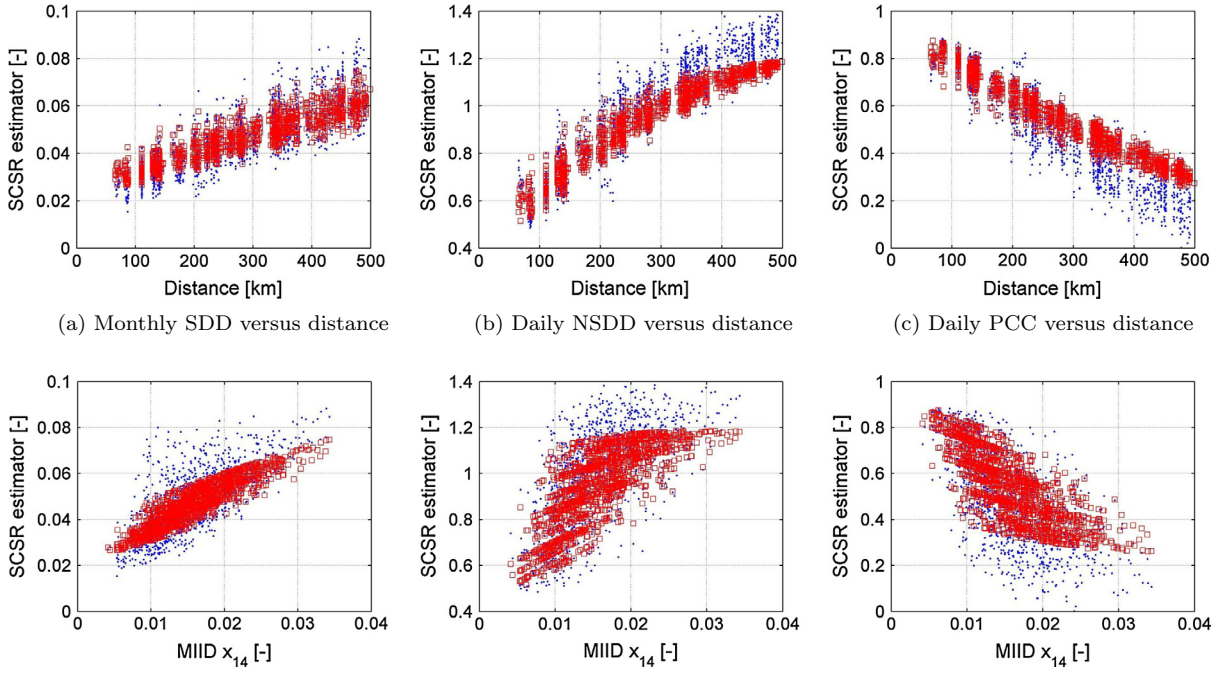
During the study the SCSR analysis was also performed for the hourly timescale by using the selected response and explanatory variables. General patterns were detected in



(a) Monthly SDD versus distance (b) Daily NSDD versus distance (c) Daily PCC versus distance

(d) Monthly SDD versus x_{14} (e) Daily NSDD versus x_{14} (f) Daily PCC versus x_{14}

Fig. 5. Predicted (red squares) and actual (blue dots) variation of the chosen response parameters with distance and MIID x_{14} (main dataset). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) Monthly SDD versus distance (b) Daily NSDD versus distance (c) Daily PCC versus distance

(d) Monthly SDD versus x_{14} (e) Daily NSDD versus x_{14} (f) Daily PCC versus x_{14}

Fig. 6. Predicted (red squares) and actual (blue dots) variation of the chosen response parameters with distance and MIID x_{14} (additional dataset). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the data for some combinations of the hourly SCSR estimators with the distance and MIID; for example, it was observed that the hourly PCC decreases with d and x_{14}. And yet, the dispersion of the data around the trend was always high, leading to the poor regression fit with the R^2 values less than 0.35. This confirms that at short timescales the correct estimation of SCSR might be impossible without taking into account local weather factors.

Table 2
Goodness of fit of the final regression models.

SCSR estimator	Model	R^2	RMSE	CV(RMSE)	MAE	MARE
Main dataset						
SDD_{month}	Eq. (6)	0.854	0.006	0.113	0.004	0.100
$NSDD_{day}$	Eq. (7)	0.852	0.080	0.088	0.062	0.076
PCC_{day}	Eq. (8)	0.846	0.068	0.120	0.053	0.107
Additional dataset						
SDD_{month}	Eq. (6)	0.733	0.007	0.158	0.006	0.135
$NSDD_{day}$	Eq. (7)	0.869	0.099	0.098	0.081	0.080
PCC_{day}	Eq. (8)	0.857	0.104	0.218	0.082	0.253

Note: The units of the goodness of fit measures are according to their definitions and may vary depending on the selected SCSR estimator.

3. Incorporating spatial correlation into univariate stochastic model

3.1. Proposed procedure

The classical univariate stochastic algorithms for synthesizing solar radiation values focus on a single location, a certain timescale and deploy the clearness index K_t as the main parameter. The calculation process involves generation of the random number sequence and its conversion to the K_t time series by using autoregressive-moving-average, Markov or other stochastic model.

As it was mentioned, a simple approach to incorporate the spatial correlation when generating the data for multiple sites is to apply it to the random number streams that drive the algorithms. The approach deploys the existing methods of linear algebra and requires the correlation matrix relating the random number streams used for individual sites C_r. In this work the same technique is adopted with the difference that the expected SCSR is evaluated by the derived regression Eqs. (6) (8), and it is used to determine C_r through iterations with the objective to match the simulated and expected values of SCSR. The calculation of C_r is associated with the identification of its functional relation to SCSR by trial and error. The relation in general is expressed as:

$$C_r = f(\mathrm{PCC}, v_i) \tag{11}$$

where v_i are the constants. PCC is chosen as a primary SCSR estimator because it makes the overall calculation procedure more robust according to the performed numerical tests.

Considering this, the proposed steps for incorporating spatial correlation into univariate stochastic algorithm are as follows:

1. Calculate the expected PCC values for the given location pairs. At a daily timescale it is done directly by using Eq. (8); at a monthly timescale Eq. (4) is employed for which the missing SDD_{month} is obtained from Eq. (6) and $\mathtt{std}(\Delta K_t)$ from the data series created by running the stochastic model for each site independently.
2. Choose (new) functional form (e.g. polynomial) and/or initial values of the constants v_i for Eq. (11).
3. Determine the correlation matrix C_r from Eq. (11) based on the expected PCC and the current values of v_i.
4. Simultaneously generate the correlated series of K_t. This only requires feeding the original stochastic algorithm at each timestep with the correlated random numbers. The latter are obtained by using the Cholesky factorization[1]:

$$U^T \times U = C_r \tag{12}$$

$$r_{corr} = r \times U \tag{13}$$

where U is the upper triangular matrix satisfying Eq. (12); and r, r_{corr} are normally distributed independent and correlated random numbers for the given locations. When a distribution other than normal is needed, r_{corr} can be modified by using inverse transform sampling.
5. Calculate from Eq. (3) the resultant (simulated) PCC for the generated K_t time series.
6. Compare the resultant and expected PCC both qualitatively and quantitatively, for example, by using scatter plot and one of the goodness of fit measures. If the deviations are not acceptable, but there is noticeable reduction compared to the previous run, update the constants v_i by fitting Eq. (11) to the current values of C_r and the resultant PCC based on the least squares method, and repeat from Step 3. If the results are poor and they differ negligibly from the previous iteration, start over from Step 2. The simulations are continued until the desired fit is achieved.

It is important to note that the exact implementation of the described procedure and its success depend on the features of selected stochastic algorithm. For example, if the algorithm performs repetitive runs until certain conditions are satisfied, the simultaneous multisite data generation (Step 4) might not be effective. The reason is that an increase in the number of locations reduces dramatically the probability of achieving the K_t values for all sites within the specified limitations. This shortcoming can be mitigated by relaxing the restrictions, but this

[1] If the correlation matrix C_r is not positive definite as required for the Cholesky decomposition, one can use the nearest positive definite matrix instead, determined by one of the existing methods of linear algebra. In this work the authors adopted the tool from D'Errico (2013).

method implies a compromise between data quality on local and regional scales.

3.2. Demonstration

The proposed procedure for incorporating SCSR into synthetic data generation was tested by using the stochastic algorithms introduced by Bohlen and Schumacher (1996) and Aguiar et al. (1988) for monthly and daily timescales, respectively. The simulations were performed for the 4 US regions and the time period (12 years) covered by the main dataset (see Section 2.2.1) and by using the corresponding average values of the monthly K_t as inputs. The total computational time on the workstation was up to 4 s per site.

The Bohlen model creates the monthly K_t time series simply by adding Gaussian noise to the long-term average values of the monthly K_t, and it is one of the few approaches known to the authors for the given timescale. The Bohlen algorithm was implemented with no adjustments.

The Aguiar model is the traditional technique which generates the daily K_t time series based on the Markov transition matrices. When applying the model, the restriction on the deviation of the generated daily K_t values from the monthly average was relaxed in order to avoid an endless loop. Even though this led to the maximum deviations of up to 9%, on the average (among all sites) the observed discrepancy was less than 3%, which was considered reasonable.

Table 3
The adopted constants for Eq. (14) at a monthly (daily) timescale.

US region	v_1	v_2
1	1.12 (1.04)	0.16 (0.06)
2	1.00 (0.98)	0 (0.04)
3	1.00 (0.98)	0 (0.06)
4	1.21 (1.04)	0.27 (0.01)

The relation in Eq. (11) was substituted by

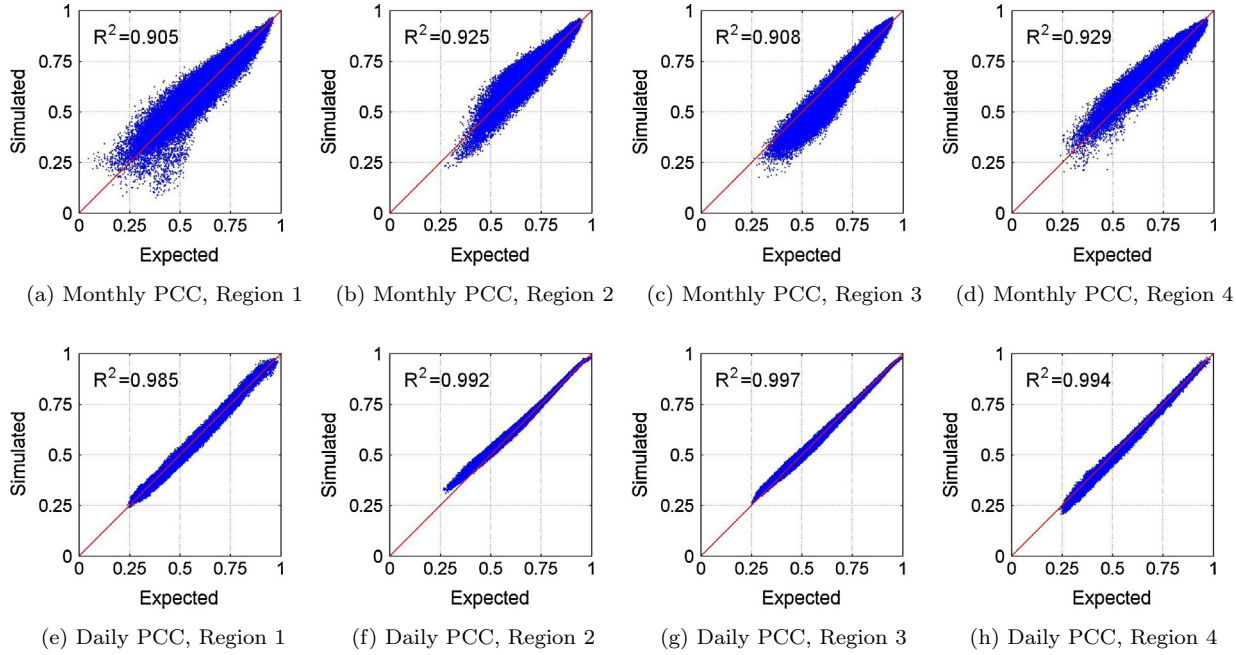
$$C_r = v_1 \times \mathrm{PCC} + v_2 \tag{14}$$

The constants were initialized as $v_1 = 1$ and $v_2 = 0$. The calculation procedure had to be repeated only once or twice, since after that no more improvements were observed in the results. The calibrated values of the constants v_1 and v_2 are given in Table 3.

The resultant PCC for the final generated synthetic data are compared to the expected (based on the regression models) values in Fig. 7. Relatively higher dispersion of the data points and thus lower R^2 at a monthly timescale is explained mainly by the shorter length of the corresponding K_t time series. Overall, the fit between the simulated and expected PCC is high ($R^2 = 0.90 - 0.99$), which confirms that it is an effective approach to enforce spatial correlation to the output of stochastic models by feeding them with the correlated random number streams.

The comparisons of the simulated and actual (based on the main dataset) PCC are presented in Fig. 8. The data fit is lower, but reasonable ($R^2 = 0.66 - 0.93$) and it differs among the selected regions with the poor results corresponding to the regions with large spatial variation in SCSR (see Fig. 2a h). The reason is that the deviations in this case include not only the error associated with incorporation of the spatial correlation into stochastic algorithm, but also the errors in the regression functions (6) and (8) and the Bohlen and Aguiar models used in the given demonstration.

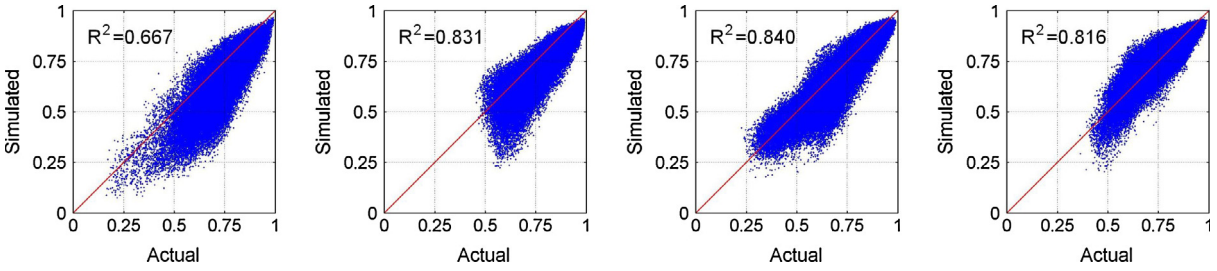
Finally, as an example, the impact of spatial correlation on the cumulative distribution of the regional (average) daily solar radiation is shown in Fig. 9. One can see that with



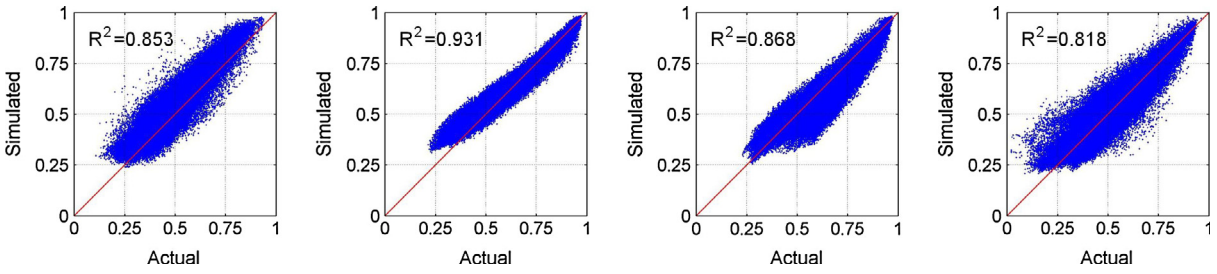
(a) Monthly PCC, Region 1 (b) Monthly PCC, Region 2 (c) Monthly PCC, Region 3 (d) Monthly PCC, Region 4

(e) Daily PCC, Region 1 (f) Daily PCC, Region 2 (g) Daily PCC, Region 3 (h) Daily PCC, Region 4

Fig. 7. Comparison of the simulated and expected PCC for the location pairs in the selected 4 US regions.

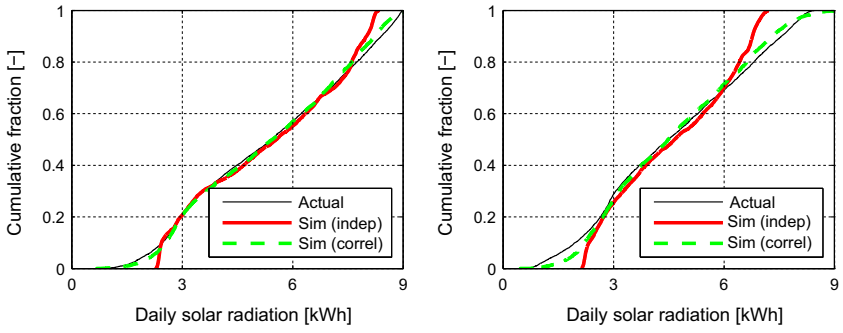


(a) Monthly PCC, Region 1 (b) Monthly PCC, Region 2 (c) Monthly PCC, Region 3 (d) Monthly PCC, Region 4

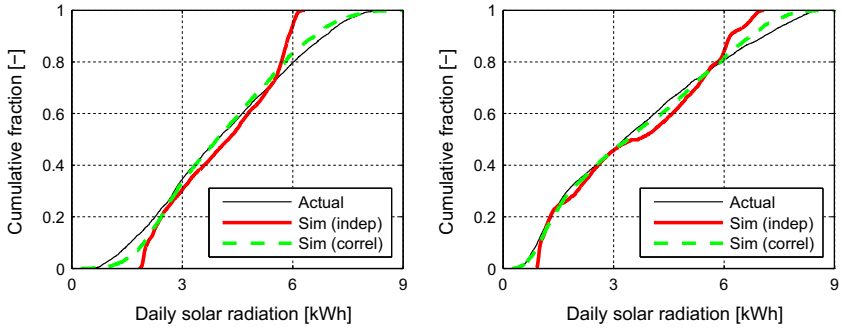


(e) Daily PCC, Region 1 (f) Daily PCC, Region 2 (g) Daily PCC, Region 3 (h) Daily PCC, Region 4

Fig. 8. Comparison of the simulated and actual PCC for the location pairs in the selected 4 US regions.



(a) Region 1 (b) Region 2



(c) Region 3 (d) Region 4

Fig. 9. Cumulative distributions of the averaged actual and simulated time series of daily solar radiation for the selected 4 US regions.

integration of SCSR the qualitative difference between the actual and synthetic regional solar radiation series reduces. The main changes are observed in the tails of the cumulative distributions: the minimum and maximum values are decreased and increased, respectively. In other words, if SCSR is ignored during the multisite generation of solar radiation data, the smoothing effect in the combined fluctuations of the solar resource is overestimated.

4. Conclusions

The presented study comprises two parts. In the first part a hypothesis was made that at long timescales general

and simple characterizations of SCSR are possible. In order to test the hypothesis, the authors performed a regression analysis of the satellite-derived monthly and daily K_t values for over 300,000 location pairs in 4 US regions. It was found that:

- The adequate estimators for the spatial correlation of solar resource are SDD_{month}, $NSDD_{day}$ and PCC_{day}.
- In addition to the distance, the relevant explanatory variable is the indicator of intersite dependence defined by the monthly average values of K_t as $x_{14} = \mathtt{std}(\Delta K^A_{t,M} - \Delta K^B_{t,M})$.
- The relation between the selected input and output parameters shows linear and quadratic trends at monthly and daily timescales, respectively.

The cross-validation of the obtained regression functions by using the additional dataset with over 1500 location pairs across Spain and Germany showed reasonable goodness of fit ($R^2 > 0.7 - 0.8$), and thus confirmed the underlying hypothesis.

In the second part, by applying the derived SCSR formulae and the existing methods of linear algebra, the authors proposed a general procedure for incorporating SCSR into univariate stochastic algorithms. The procedure deploys the common technique of enforcing spatial correlation between output parameters by feeding a given stochastic model with the spatially correlated random number streams. The numerical tests were performed by using two conventional stochastic solar radiation algorithms of different complexities. In both cases a good match was observed between the expected (from the regression models) and simulated values of the spatial correlation, which confirmed the effectiveness of the proposed procedure. The comparisons of the generated and actual solar radiation values also demonstrated that the quality of the synthetic data is reasonable and it improves with integration of SCSR.

References

Aguado, E., 1986. Local scale variability of daily solar radiation San Diego County, California. J. Clim. Appl. Meteorol. 25 (5), 672 678.

Aguiar, R., Collares Pereira, M., Conde, J., 1988. Simple procedure for generating sequences of daily radiation values using a library of Markov transition matrices. Solar Energy 40 (3), 269 279.

ASDC, 2013. NASA Atmospheric Science Data Center. Surface Meteorology and Solar Energy. <http://eosweb.larc.nasa.gov/sse> (accessed 21.11.14).

Badescu, V., 2008. Modeling Solar Radiation at the Earth's Surface: Recent Advances. Springer.

Badosa, J., Haeffelin, M., Chepfer, H., 2013. Scales of spatial and temporal variation of solar irradiance on Reunion tropical island. Solar Energy 88 (0), 42 56.

Bohlen, M., Schumacher, J., 1996. Time series analysis of the long term monthly horizontal solar radiation. In: Proceedings of EuroSun'96, pp. 1503 1507.

D'Errico, J., 2011. Matlab Functions for Polynomial Regression Modeling. <http://www.mathworks.com/matlabcentral/fileexchange/34765 polyfitn> (accessed 21.11.14).

D'Errico, J., 2013. Matlab Function for Finding the Nearest Positive Definite Matrix. <http://www.mathworks.com/matlabcentral/fileexchange/42885 nearestspd> (accessed 21.11.14).

Duffie, J., Beckman, W., 2006. Solar Engineering of Thermal Processes. John Wiley & Sons.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157 1182.

Hansen, B.E., 2013. Econometrics. Draft Graduate Textbook. University of Wisconsin. <http://www.ssc.wisc.edu/bhansen/econometrics/> (accessed 21.11.14).

Hoff, T., Perez, R., 2012. Modeling PV fleet output variability. Solar Energy 86 (8), 2177 2189.

Khalili, M., Brissette, F., Leconte, R., 2009. Stochastic multi site generation of daily weather data. Stoch. Env. Res. Risk Assess. 23, 837 849.

Koller, D., Sahami, M., 1996. Toward optimal feature selection. In: Proceedings of 13th Intl. conf. on Machine Learning, pp. 284 292.

Lin, M., Lucas, H.C., Shmueli, G., 2013. Research commentary too big to fail: large samples and the p value problem. Inform. Syst. Res. 24 (4), 906 917.

Meteonorm, 2013. Handbook Part II: Theory. Version 7.0.20. <http://meteonorm.com> (accessed 21.11.14).

NSRDB, 2013. National Climatic Data Center. National Solar Radiation Database. <http://www.ncdc.noaa.gov/data access/land based station data/land based datasets/solar radiation> (accessed 21.11.14).

Perez, R., Ineichen, P., Seals, R., Zelenka, A., 1990. Making full use of the clearness index for parameterizing hourly insolation conditions. Solar Energy 45 (2), 111 114.

Sinnott, R.W., 1984. Virtues of the haversine. Sky Telescope 68, 159.

Suckling, P., 1995. Spatial coherence of incoming solar radiation for the Southern Piedmont. Southeast. Geogr. 35 (2), 183 193.

Watgen, 1992. User's Manual. Watsun Simulation Laboratory, University of Waterloo.

Wilks, D., 1999. Simultaneous stochastic simulation of daily precipitation, temperature and solar radiation at multiple sites in complex terrain. Agr. Forest Meteorol. 96 (1), 85 101.

Wilks, D., Wilby, R., 1999. The weather generation game: a review of stochastic weather models. Prog. Phys. Geogr. 23 (3), 329 357.